图书情报工作
第 62 卷 第 15 期 2018 年 8 月

基于词向量的微博情感倾向分类研究*

■ 刘勘 袁蕴英
中南财经政法大学信息与安全工程学院 武汉 430074

摘要:[**目的/意义**]微博已成为大众情感表达的重要平台,微博的情感分析在舆情分析、用户体验、商机挖掘等方面有着重要的作用。[**方法/过程**]提出的情感倾向分类算法 WE_SDAE 使用单词嵌入的方式将微博表示成一个低维稠密向量,然后通过添加正则项和加噪处理的方式将基本的自动编码器算法优化成深层噪音自动编码器,并在顶层添加分类器,实现情感倾向分类。考虑到微博用词灵活,还从单字和词语两个粒度训练模型。[**结果/结论**]实验结果表明,基于单字粒度的模型表现优于基于词语粒度的模型。此外,对比实验显示 WE_SDAE 算法优于传统的 SVM、Naive-Bayes、XgBoost 等相关算法;单词嵌入的方式优于传统的向量空间模型表示方法,能在微博情感分析中取得较好的效果。

关键词:情感分析 分类 自动编码器 微博
分类号:TP391
DOI:10.13266/j.issn.0252-3116.2018.15.011

微博包含了用户丰富的情感信息。随着微博的全民普及,越来越多的用户习惯于在微博上描述个人的生活经历,表达自己的情感体验或者点评社会时事热点。微博中往往记录着每位用户的点滴喜怒哀乐。这些情感信息的提取与研究能够帮助政府开展更为及时有效的舆情引导,帮助企业改进产品的用户体验,帮助创业者挖掘潜力巨大的商机。但是,微博文本较短,表达不规范,传统的方法已经无法很好地满足微博的处理需求,迫切地需要可以高效提取微博情感倾向的新方法。

1 相关研究

针对情感分析的研究主要是从粗粒度和细粒度两个方面展开。粗粒度主要是指篇章和句子层面,关注于整篇文档或整个句子积极或消极的情感态度;细粒度则主要关注字词层面,关注整体情感下面的细节态度,如积极情感下的高兴、漂亮、点赞、轻松等更具体的情感态度。由于微博篇幅较短、字数不多,难以深入到细粒度研究,而且大多数微博表达态度明确单一,因此适合于粗粒度的情感分析。

目前,微博领域的情感分析主要包括基于情绪知识和基于传统机器学习方法两类。情绪知识可以是特定的情绪词、标签信息、情感词典、表情符号等。P. D. Turney 等[1]选取"excellent"和"poor"两个情绪词作为种子词,然后抽取句子中的形容词,与这两个种子词分别计算互信息,按照互信息的值确定句子的情感倾向;任远等[2]针对情绪词词性提出了更为精细的划分方法,并尝试引入主题因素,设计了面向主题的自适应情感分类方法;L. Barbosa 等[3]考虑了 twitter 的结构特征和词汇信息,利用了极性标签信息完成对 twitter 情感倾向性的判断;庞磊[4]等人利用微博中的情绪词和表情图片来获取微博的情感倾向;潘明慧等[5]将表情符词典与传统的情绪词典进行结合,并制定了否定语法规则,识别出微博中表达的喜、哀、怒、惧、恶、惊 6 种情绪,提高了微博情绪倾向分析的精度;刘全超等[6]构建了情感分析用词词典、网络用语词典和表情符号库,并结合微博间的转发和评论关系来设计微博情感倾向性判定算法。

还有采用机器学习模型来解决情感倾向分析问题。A. Bakliwal 等[7]整合了语义特征和 twitter 相关特征,使用 SVM 将 twitter 划分成正面、负面和中性三类;B. Johan 等[8]利用心理测量工具抽取六维情感向量来

* 本文系国家社会科学基金项目"基于文本挖掘的网络谣言预判研究"(项目编号:14BXW033)研究成果之一。
作者简介:刘勘(ORCID: 0000-0002-9686-9768),教授,博士,E-mail:liukan@zuel.edu.cn;袁蕴英(ORCID: 0000-0003-1713-1624),硕士研究生。
收稿日期:2017-12-24 修回日期:2018-04-22 本文起止页码:92-101 本文责任编辑:徐健

完成微博情感倾向性的判断；C. Tan 等[9]假定存在社会关系的用户更有可能拥有相似的观点，因而提出将社会关系信息引入到传统的 SVM 中来构建情感倾向性模型；刘志明等[10]分别使用 SVM、朴素贝叶斯和 N-Gram 方法进行情感极性的分类；朱玺等[11]优化了半监督学习方法 reserved self-training 的特征选择法方法和迭代终止条件，有效防止了过拟合现象的产生，提升了模型的准确率；孙建旺等[12]抽取微博中的动词和形容词作为特征，依据层级结构来完成特征降维，通过表情符号计算特征极值，最后借助 SVM 将微博文本划分成正面、负面和中性三类。

总的来看，对于基于情绪知识的方法来说，构建情绪知识体系会带来较高的人工成本，无论是情绪词、标签还是情感词典，建立的时候都需要进行人工标注。此外，基于情绪知识的方法的使用范围较为受限，只能处理包含这些情绪知识的微博。考虑到中文中广泛存在的多义性，同一个情绪知识在不同的上下文中往往还会表达出截然不同的情感倾向。对于基于传统机器学习模型的方法而言，现有的方法基本都围绕着 SVM 展开，采取的方式集中于寻求特征上的扩充与完善，对分类模型本身的贡献较少。而自动编码器算法具有强大的非线性学习能力，在自然语言处理的很多方面都取得了很好的效果，是一个值得尝试的方向。此外，机器学习模型的大部分特征都来自于向量空间模型。一方面向量空间模型无法描述出词与词之间共享的语义信息[13]，会影响其情感倾向性判定的准确性；另一方面，微博口语化程度高，缩写频繁、搭配随意，通过向量空间模型转化成向量后具有高维稀疏的特点[14]，由此带来的维数灾难问题也一直困扰着传统的机器学习方法。而单词嵌入可以在有限维度上较好地刻画出每个词的语义特征，对解决上述问题很有帮助。

2 基本思路

本文不依靠任何人工标注的情绪知识，结合微博本身高维稀疏的特点，提出了一种基于单词嵌入的微博向量和深层噪音自动编码器的微博情感倾向分类算法 WE_SDAE(Word Embedding_Stacked Denoising Auto-Encoder)。算法主要包括三个核心步骤：使用单词嵌入获取词向量；改进自动编码器提取低维抽象特征和有监督全局调整参数以完成情感倾向分类。

首先，经过预处理后的微博文本通过单词嵌入(word embedding)的方法表示成一个分布式低维稠密向量。其次，这些向量将被输入到优化后的深层噪音自动编码器中，经过逐层无监督的非线性学习，转化成抽象特征。自动编码器的优化过程包含两个部分，分别是往目标函数添加 L1 正则项和在预处理后的训练数据中添加噪音。最后，在自动编码器的顶层增加一个分类器，进行有监督的训练，全局调整参数，得到最终的情感倾向分类算法。考虑到微博中存在大量的缩写与简写，用词灵活，传统中文分词工具的切分可能会带来较多信息的损失，算法从单字和词语两个不同的粒度进行了处理。情感倾向分类算法的流程如图 1 所示：

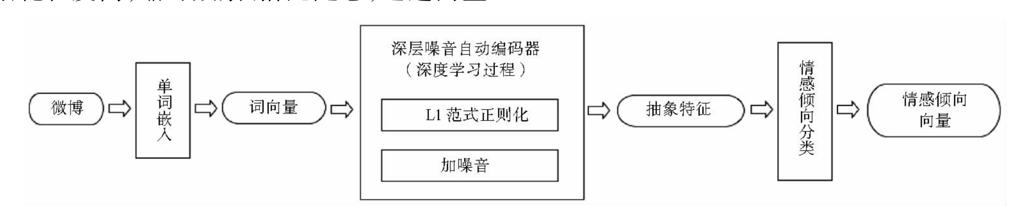


图 1 WE_SDAE 算法流程

3 词向量获取

3.1 单词嵌入

单词嵌入的实现方法有很多种，目前最为流行的是 CBOW(Continuous Bag-Of-Word)[15]和 Skip-gram[16]两种方法。在语义相关的工作中，Skip-gram 的表现优于 CBOW，因而采用 Skip-gram 方法。

当给定一组训练单词 $w(t-1)$，$w(t-1)$，$w(t)$，$w(t+1)$，$w(t+2)$时，Skip-gram 将目标单词 $w(t)$作为输入，经过映射层的处理后，输出目标单词所在上下文的单词，即基于相似的单词拥有相似语境的基本假设，试图通过当前目标单词来预测其语境信息。算法的流程见图 2。

利用最大似然函数的思想，该概率语言模型的目标函数如公式(1)所示：

$$p(Context(w)|w)=\prod_{u\in Context(w)}p(u|w) \quad \text{公式(1)}$$

当使用 Hierarchical Softmax 框架求解该语言模型时，输入层是中心词 w 的词向量 v_w，输出层是 $Context(w)$对应的 Huffman 树，其叶子结点是上下文语境中的

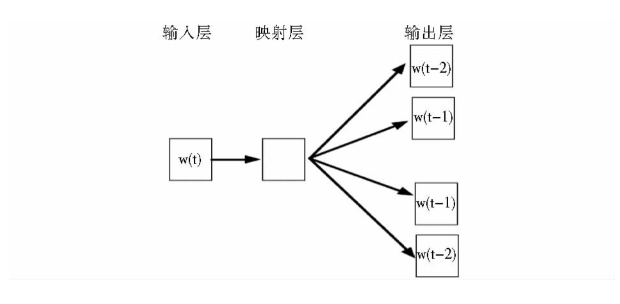


图2 Skip-gram 结构

词，权值是各词在语料中出现的次数，此处的映射层是一个恒等映射，保留的原因仅是为了与上文结构一致，如图3所示：

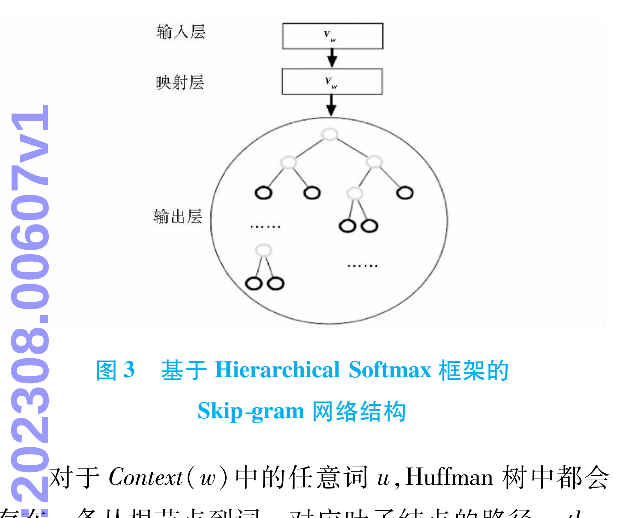


图3 基于 Hierarchical Softmax 框架的 Skip-gram 网络结构

对于 $Context(w)$ 中的任意词 u，Huffman 树中都会存在一条从根节点到词 u 对应叶子结点的路径 $path_u$，$path_u$ 中共包含 l_u 个结点，其中第 j 个结点对应的编码是 $d_{u,j}\in\{0,1\}$，则该词 u 的 Huffman 编码可以表示成 $[d_{u,2},d_{u,3},\cdots,d_{u,l_u-1}]$（根结点不对应编码）。$path_u$ 中第 j 个非叶子结点对应的向量是 $w_{u,j}$，$path_u$ 上共存在 l_u-1 个非叶子结点，每个叶子结点都可以看成是一个分支，对应一个二分类，产生一个概率，将这些概率相乘就是 $p(u|w)$，如公式(2)所示：

$$p(u|w)=\prod_{j=2}^{l_u}p(d_{u,j}|v_w,\psi_{u,j-1}) \qquad \text{公式(2)}$$

其中，v_w 是中心词 w 的词向量，将被随机初始化，然后通过随机梯度下降算法迭代优化，得到最终的结果，即包含原词语义信息的词向量。

3.2 微博向量

假设一条微博 S 是由单词 $w_1,w_2,w_3,\cdots,w_n$ 组成，每个单词 $w_i(1\leqslant i\leqslant n)$ 都可以通过单词嵌入方法 Skip-gram 获得一个词向量 v_w。根据 T. Mikolov 的研究成果[16,17]，通过 Skip-gram 获取的词向量之间的基本算术运算具有丰富的潜在语义信息。比如将单词“德国”的词向量与单词“首都”的词向量进行相加，获取的词向量与单词“柏林”的词向量十分相似，又比如单词“国王”的词向量减去单词“男”的词向量，再加上单词“女”的词向量，得到的词向量接近于单词“女王”的词向量。依据 Skip-gram 的这个特性，将微博 s 中所有单词的词向量的平均数作为该微博的词向量，如公式(3)，思路如图4所示。

$$v_s=\frac{1}{n}\sum_{i=1}^{n}v_{w_i} \qquad \text{公式(3)}$$

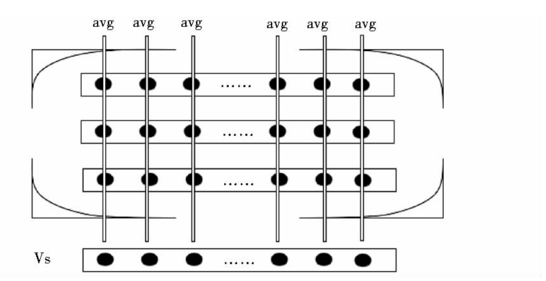


图4 计算单条微博的词向量

所有微博数据按照上述方法处理后可以得到一个微博矩阵 S，公式(4)如下：

$$S=\begin{matrix} s_1^T \\ s_2^T \\ \cdots \\ s_M^T \end{matrix} \in R^{m\cdot d} \qquad \text{公式(4)}$$

其中，M 代表数据中微博的条数，d 代表向量的维数。

3.3 词粒度和字粒度

首先利用中文分词工具对微博文本进行切分，将划分之后得到的词语作为每条微博的基本组成单元，即前文微博 S 中的词 $w_1,w_2,w_3,\cdots w_n$ 是指分词后的词语。比如“中国加油”将划分成“中国/加油”。本文把这种方式看作基于词粒度的词向量获取方法。

微博是典型的短文本，上下文信息有限、缩写频繁、噪声大，分词工具的分词结果并不准确，往往存在较多歧义。比如“这是一个高大上网站”中的“高大上网站”如果使用传统的分词技术，得到的结果为“高大/上/网站”或者“高大/上网/站”，无法体现出微博想要表达的正确语义。此外，微博中不断出现大量新词。这些词可能是网友最新创造出来的，比如“城会玩”“活久见”等，也可能是原词，但已在微博上引申出了新的含义，比如“我来安利一下这个 app”中“安利”不再代表某品牌，而是表示“强烈推荐”的意思。这些信息都可能在分词过程中损失，导致微博的分词结果无法令人满意。因此，本文借鉴了 X. Zheng 等[18]在解决词性标注问题时的方案，提出了基于字粒度的词向量获取方法，直接将微博中的所有字都拆分开，把单个

字作为微博的基本组成单元，即前文微博 S 中的单词 $w_1,w_2,w_3,\cdots,w_n$ 将代表中文中的单个字。比如“高大上网站”将被表示成“高/大/上/网/站”。

4 自动编码器提取特征

自动编码器是深度学习中一种重要的训练模型，一直以来在自然语言处理中取得了较好的效果。接下来将利用自动编码器学习文本特征，并在此基础上添加正则项以约束算法的学习能力，对输入数据进行加噪处理以提高鲁棒性，叠加多个自动编码器以提高特征抽象能力。

4.1 L1 范式正则化

自动编码器强大的非线性表达能力虽然有助于获取抽象特征，但容易出现过拟合的问题，即对个体所特有的信息也进行了充分的学习。不同微博的差异较大，不可避免地包含大量特有特征。如果直接采用基本的自动编码器算法，抽取的特征向量很可能无法反映出数据的本质共性，训练得到的模型的泛化能力特别差，无法进行有效的推广扩展。因此需要对自动编码器的学习能力进行了约束。

L1 范式正则化是一种常用的变量选择方法，被广泛运用于各种算法的改进工作。把自动编码器系数的绝对值函数当作惩罚项，压缩系数值，将绝对值较小的系数直接压缩为0，从而保证算法参数的稀疏性，避免过分学习微博中的非显著特性。具体计算如公式(5)和公式(6)所示：

$$L(x,z)=KL(x\parallel z)+Lasso(\theta) \qquad \text{公式(5)}$$

$$\mathrm{Lasso}(\theta)=\lambda\sum_{j=1}^{|\theta|}|\theta_j| \qquad \text{公式(6)}$$

其中，损失函数为 Kullback-Leibler 散度，使用经典的随机梯度下降算法进行训练，λ 是 L1 范式的参数，值越大，惩罚力度就越大，训练得到的结果越稀疏，具体取值需要根据实际数据进行调试，以均衡算法的拟合能力和泛化能力。

4.2 加噪处理

考虑到微博输入的随意性很高，大量网民在发布微博时都会使用缩写和简写，甚至一些个性化的语言和符号，同时由于输入时较为匆忙，也经常会出现多输、漏输甚至错输文字的现象，这要求面向微博的情感倾向分类算法必须具有较强的鲁棒性。

针对这些问题，可以在微博的词向量中添加一定量的噪音，增加训练数据的干扰性。P. Vincent 等[19]添加噪音的方式是随机选择一定比例的数据强制变为0，本文则除了会选取部分数据强制变为0，还会挑选一定比例的数据强制变为标准正态分布中的一个随机数。前者考虑到输入向量中的数据缺失情况，训练得到的自动编码器应该具备还原这些缺失值的能力；后者考虑到微博输入中广泛存在的不规范性，保证了算法能够避免受到这些个性化或者无关输入的干扰。

向量 x 输入编码器后，通过线性变化，再经过激活函数的处理后得到编码结果 y，计算见公式(7)。编码结果 y 又会输入到解码器中，经处理得到重构后向量 z，计算如公式(8)所示：

$$y=f_\theta(x)=s(Wx+b) \qquad \text{公式(7)}$$

$$z=g_{\theta'}(y)=s(W'y+b') \qquad \text{公式(8)}$$

编码的参数是 $\theta=\{w,b\}$，解码的参数是 $\theta'\{W',b'\}$。其中，W 是一个 $d'\times d$ 的权重矩阵，W'是 W 的转置，即 $W'=W^T$，b 和 b'是相应的偏倚向量。优化目标是使重构后的向量 z 尽量接近输入向量 x，即最小化重构带来的损失，得到最优参数 θ^* 和 θ'^*，如公式(9)所示：

$$\theta^*,\theta'^*=\underset{\theta,\theta'}{\mathrm{argmin}}L(x,z)=\underset{\theta,\theta'}{\mathrm{argmin}}L(x,g_{\theta'}(f_\theta(x))) \qquad \text{公式(9)}$$

输入向量 x 在加入噪音后变成 $\tilde{x}$，调整后利用随机梯度下降算法得到编码和解码的最优参数 θ^* 和 θ'^*，如式(10)所示：

$$\theta^*,\theta'^*=\underset{\theta,\theta'}{\mathrm{argmin}}L(x,z)=\underset{\theta,\theta'}{\mathrm{argmin}}L(x,g_{\theta'}(f_\theta(\tilde{x}))) \qquad \text{公式(10)}$$

4.3 深度自动编码器

将多个噪音自动编码器进行叠加后就形成了深度学习网络。叠加的自动编码器层数越多，学习抽象特征的能力就会越强[20]。在训练的过程中，K-1 层自动编码器输出的抽象特征向量加上噪音后作为 K 层自动编码器的输入向量，K 层自动编码器通过最小化损失函数，使得解码器处理后的重构向量与未加噪音的原始输入向量尽量接近。不断优化调整参数，获得最优解后，K 层自动编码器将丢弃解码器部分，并把编码器处理后的抽象特征向量加上噪音作为 K＋1 层的输入向量，继续进行下一层的训练。如此循环，逐层训练，就形成了深层噪音自动编码器模型，最终得到特征向量。其结构图见图5。

5 情感倾向分类

前文的深层噪音自动编码器只是通过无监督的方式提取了微博文本中的抽象特征，还无法完成情感倾

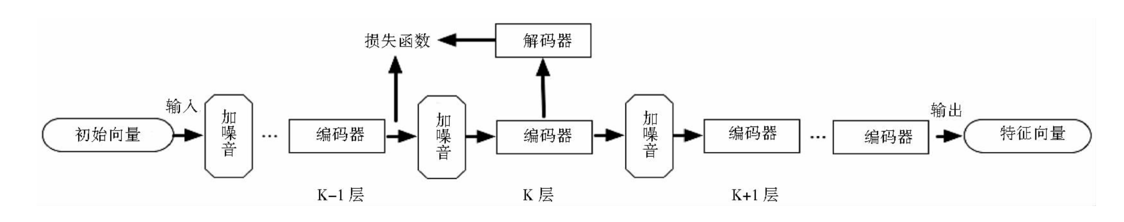


图 5　深层自动编码器结构

向性分类工作，需要在深层噪音自动编码器的最后一层的后面再添加一层分类器，作为最终的输出层。

该层的输入向量是最后一层自动编码器的输出结果，输出向量是微博情感倾向向量。此处定义情感标注集 T = {正面，中性，负面}，因而微博情感倾向向量的维数是 3，形如 $v_e = [a, b, c]^T$，其中 a 代表该微博正面情感标注，b 代表情感中性标注，c 代表负面情感标注，值为 1 表明存在情感，值为 0 表明不存在情感。假设一条微博仅表达一种主要情感，比如 $v_i = [1, 0, 0]^T$ 代表该微博 i 的情感倾向是正面的。

该层的处理过程包括线性变换输入向量和激活函数非线性调整输出最终向量两个步骤。这里选用广泛应用于多分类问题的 softmax 函数作为激活函数。计算如公式（11）所示：

$$f(x) = softmax(wx + b) \qquad \text{公式(11)}$$

其中，$w \in R^{h \times d}$（h 是情感类别数量，此处为 3，d 是输入向量的维数）代表权重矩阵，$b \in R^h$ 代表偏置项。该层仍然选用 KL 散度作为损失函数，衡量了算法输出向量与目标向量之间分布的相似性，计算公式如（12）所示：

$$L(X, f(X)) - KL(x \parallel f(X)) \qquad \text{公式(12)}$$

至此，整个微博情感倾向分类算法 WE_SDAE 的训练过程可以总结为两个步骤。首先，算法中的单层噪音自动编码器将依次进行无监督学习，不断抽象迭代，逐步从原始输入向量中提取得到数据的本质特征，作为后续训练的基础。这一步中每一层的训练过程都是相对独立的。然后借助已经标注好的情感倾向向量，将进行全局的有监督学习，训练方式仍然采用标准的梯度下降法即可。该优化过程不仅调整顶层分类器中的参数，也将对前面已经训练好的所有噪音自动编码器中的参数进行微调，保证了整个过程具有最佳的学习能力，见图 6。

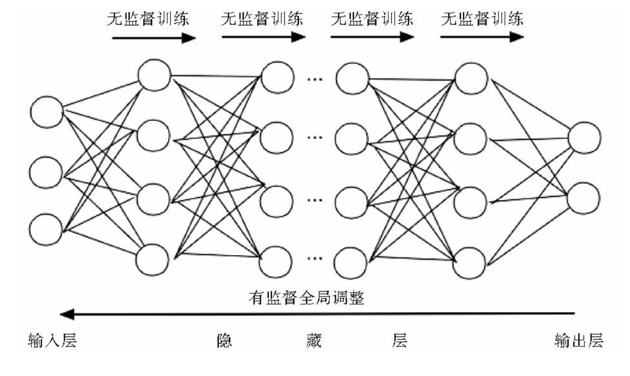


图 6　WE_SDAE 算法训练过程

6　实验结果与分析

6.1　数据源选择及数据预处理

实验数据来自 2014 年 NLPCC 会议的 COAE（Chinese Opinion Analysis Evaluation）微博测评数据集。该数据集共包含了 5 000 条已标注好的微博数据，其中 2 656条带有正面情感倾向，2 344 条带有负面情感倾向，在此基础上又另外爬取和人工标了 2 500 条中性情感倾向的微博数据，得到最终 7 500 条实验数据，具体见表 1，数据样例见图 7。同时数据按照 4∶1 的比例随机被拆分成两部分，分别作为训练集和测试集。

表 1　实验数据

类别	正面（条）	中性（条）	负面（条）
数据量	2 656	2 500	2 344


```
<weibo id="227" emotion-type1="like" emotion-type2="none">
  <sentence id="1" opinionated="Y" emotion-1-type="like" emotion-2-type="none" keyexpression1="好幸福">
到站了，广播响起「请带好你的贵重物品下车」然后一个男生对他女朋友说「走啦，贵重物品！」这个细节，真是好幸福。。。。<
</weibo>
<weibo id="228" emotion-type1="happiness" emotion-type2="none">
  <sentence id="1" opinionated="Y" emotion-1-type="happiness" emotion-2-type="none" keyexpression1="很高兴">
自慧心理:网络用语等级考试阅读题：亲耐的童鞋，盆友们，很高兴我们在国脖里相遇。</sentence>
  </sentence>
</weibo>
<weibo id="229" emotion-type1="disgust" emotion-type2="none">
  <sentence id="1" opinionated="N" emotion-1-type="disgust" emotion-2-type="none" keyexpression1="不伦不类">
突然感觉新浪微博的聊天客户端有点不伦不类……我真的有要用客户端的需求吗？</sentence>
</weibo>
<weibo id="230" emotion-type1="happiness" emotion-type2="none">
```


图 7　标注好的微博情感样例数据

COAE2014 中的微博数据集已经经过了一定的清洗整理，去除了表情符号和系统自动生成的“转发微博”等信息，保存成了文本格式。在此基础上，又进行了如下 4 步预处理：

（1）拆分多次转发的微博。比如数据集中的某条微博为“看看相片，看看孩子可怜的身体，愤怒！！//@张蜀梅：是真的吗？//@阿子：他是从罗马尼亚流窜过来的么??”，这实际包含了多位用户的多条微博，因而需要分割还原。

刘勘，袁蕴英. 基于词向量的微博情感倾向分类研究[J]. 图书情报工作,2018,62(15):92-101.

（2）去除无关文本，比如“回复@ fyx 璇：哈哈，我可是鲁能泰山的忠实球迷，今年中超南京客场我可找你啊”中“回复@ fyx 璇：”部分并不是用户发布的微博内容，与情感倾向分析无关。

（3）去除了链接的地址内容，但保留关键词“http”。微博中的链接都是系统自动生成的短链，地址本身并不包含太多信息，但是考虑到添加链接的行为可能表明用户对自己立场的坚定，因而表明添加行为的信息还是需要保留。比如“质量很垃圾！！我在：http://t.cn/zjOsElS”。

（4）去除@的名字内容，但保留关键词“@”。@行为一般发生在作者带有正面或者负面情感倾向时，但是名字内容与其情感无关，因而可以直接去除。比如“雷克萨斯不错，不过日本车，哎！还是喜欢奥迪！@Nickole 星 @夏日料理王 @ss628 @咘噜咘噜咘噜的蓝 @潘炜晨”。

除此之外，在预处理过程中强调保留了标点符号和话题标签信息。这是因为标点符号是用户情感的重要表现载体，不同标点符号倾向的情感信息往往也是不同的，比如“？”表示疑问，经常在负面情感倾向中使用[21]。话题标签是微博讨论内容的精简概括，往往也带有特定的情感信息，有助于更好地判定情感倾向。比如“#信用卡里传来了绝望的呜咽#”等。

6.2 评价指标

使用平均 F 值 F_{avg} 来筛选参数，评价实际效果时则具体分析了每类情感倾向 C_i 的 F_i 值[22-23]。F_{avg} 值是各类情感倾向 F_i 值的平均值，计算公式如(13)所示：

$$F_{avg} = \frac{\sum_{i=1}^{C} F_i}{C} \qquad \text{公式(13)}$$

其中，C 是不同情感倾向类别的个数，在本文中 C 取值为 3。

F_i 值综合考虑了情感倾向类 C_i 的正确率与召回率。其值越大，代表情感倾向分类在该类上的表现越好。计算公式如(14)和(15)所示：

$$F_i = \frac{2 \cdot precision_i \cdot recall_i}{(precision_i + recall_i)} \qquad \text{公式(14)}$$

$$Frecision_i = \frac{m_{right_i}}{m_{right_i} + m_{wrong_i}}, \quad recall_i = \frac{m_{right_i}}{m_{all_i}} \qquad \text{公式(15)}$$

其中，m_{right_i} 是被正确分到情感倾向类 C_i 的微博数量，m_{wrong_i} 是被错误分到情感倾向类 C_i 的微博数量，m_{all_i} 是情感倾向类 C_i 中实际包含的微博数量。

6.3 算法实现

本文提出的情感倾向分类算法 WE_SDAE 在具体实现时可以分成两个步骤，分别是获取微博向量和情感分类，同时考虑了字和词两种不同的粒度。

（1）获取微博向量。预处理之后的微博首先会被切分，然后通过单词嵌入的方式获取每个字或者词的向量，最后通过公式计算得到最终的向量。

在切分微博数据时，字粒度的处理比较简单，直接将清洗好的微博按照单个字进行一一分割。比如“喜欢这款产品！”分割后得到“喜/欢/这/款/产/品/！”。当粒度要求为词时，使用中文自然语言处理中最常使用的 NLPIR 汉语分词系统对微博数据进行划分。“喜欢这款产品！”的划分结果将会变成“喜欢/这/款/产品/！”。处理过程中需要注意的是每个标点符号也作为一个单独的字或者词保留在划分后的数据中。

训练词向量的 Skip-gram 方法已经在开源工具 word2vec 中得到了高效的实现，实验中直接使用该工具获取微博数据的词向量。根据 V. Mikolov 等[24]的观点，训练时使用的数据量越大，训练得到的词向量就能更好地描述该词的语义特征，因而将使用 COAE2014 的全量 40 000 条微博数据训练每个词向量。在粒度为字的词向量训练过程中，微博的基本构成单元是单个字，word2vec 将为每个字训练得到一个对应的词向量；在粒度为词的词向量训练过程中，微博的基本构成单元变成了分词后的词语，word2vec 将会为每个切分后的词语训练得到一个对应的词向量。为了保证基于字粒度的模型和基于词粒度的模型具有更强的可比性，在训练词向量时将采用同一组参数，参数内容如表 2 所示：

表 2 Word2vec 参数列示

参数名称	参数值
词向量维数(size)	300
上下文窗口(window)	8
最低频率(min-count)	5
采样阈值(sample)	1e-3
迭代次数(iter)	10

根据中文词向量方面的研究经验，将词向量的长度设为 300，即每个词都会被映射到维数为 300 的连续空间中。上下文窗口设置为 8，即上下文语境信息将由该词前面 8 个词和后面 8 个词组成。最低频率为 5 保证了只有在数据集中出现次数大于 5 的词才会被加入到词向量库中，那些未被加入的词将会被随机初始化。采样阈值 1e-3 决定一个词在训练过程中是否会

被采样，如果一个词出现的频率超过了该阈值，该词就会被采样以节约训练时间。本次词向量的迭代次数为 10 次。

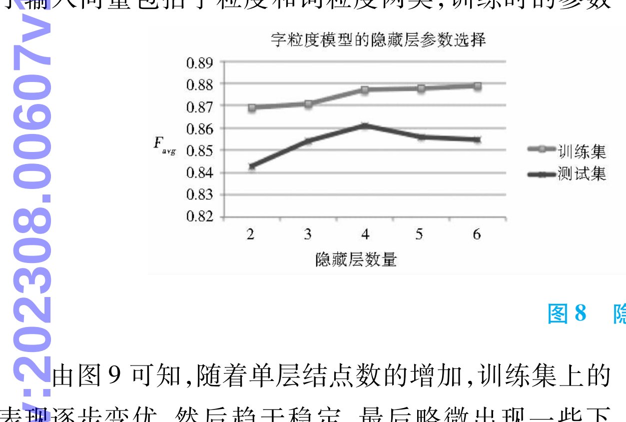
经过上述步骤，无论字粒度还是词粒度的数据都已得到了对应的词向量库，接下来将依次处理得到微博向量。对于处理后的字粒度数据，每条微博中的每个字对应一个词向量，汇总这些向量，根据算法设计中的公式(3)可以计算得到该微博的向量；对于处理后的词粒度数据，汇总向量来自于词粒度的词向量库，同样按照公式(3)计算获得最终结果。

(2)实现分类算法。利用深度学习库 theano 实现 WE_SDAE 算法。训练过程中的关键参数包括隐藏层数量、单层结点个数、正则化惩罚系数和加噪比例。由于输入向量包括字粒度和词粒度两类，训练时的参数选择也将分开进行。候选的参数值如表 3 所示：

表 3　候选参数值

参数名称	参数候选值
隐藏层数量	2/3/4/5/6
单层结点数量	100/300/500/700
正则化惩罚系数	1e-1/1e-2/1e-3/1e-4/1e-5
加噪比例	0.1/0.2/0.3/0.4/0.5

由图 8 可知，在训练集上的整体表现是随着隐藏层数量的增加而越来越好，说明隐藏层越多，从训练集中提取的抽象特征包含的信息就越丰富，但也可能出现过拟合。从测试集上的效果看，表现呈单波峰状，基于字粒度的模型在层数为 4 时达到最优，而基于词粒度的模型在层数为 3 时就已经达到了最优。



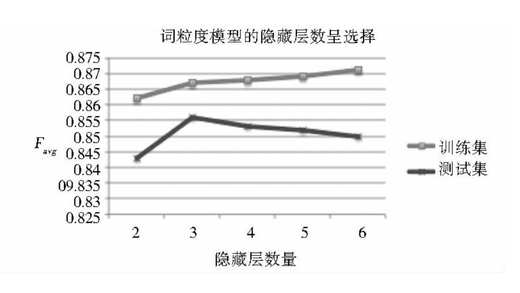
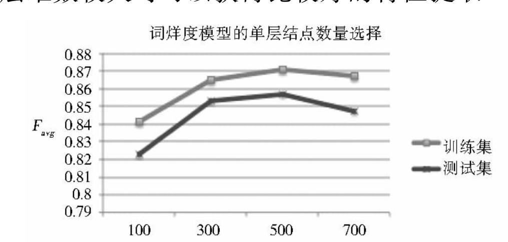


图 8　隐藏层选择

由图 9 可知，随着单层结点数的增加，训练集上的表现逐步变优，然后趋于稳定，最后略微出现一些下滑。这也符合 Y. Bengio 之前的研究成果[25]，自动编码器在输出层维数较大时可以获得比较好的特征提取效果。算法在测试集上的表现仍然是单波峰状，而且无论是字粒度模型还是词粒度模型均在单层结点数量为 500 时达到最佳。



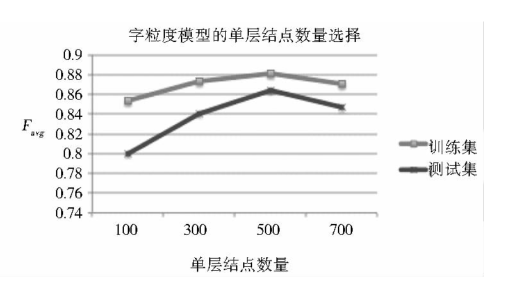


图 9　单层结点数量选择

由图 10 可知，正则化惩罚系数越大，越多的参数会被置为 0，学习能力将被进行更强的限制。在实验中，随着正则化惩罚系数的不断变小，在训练集和测试集上的表现从整体趋势来看都是先变好，后变差，并且训练集和测试集上的表现差异也是在同一个值处达到最小。无论是字粒度模型还是词粒度模型，最佳正则化惩罚系数都是 1e-3。

由图 11 可知，随着加噪比例的增加，训练集和测试集上的表现差异越来越小，并且在测试集上的表现越来越好，说明对于微博这类用词较为随意的短文本，增加噪音的确增强了算法的抗干扰能力。整体上，词粒度模型的波动比字粒度模型更为明显，说明划分成词的过程中出现了一些偏差，存在更多的噪音信息。词粒度模型的最优加噪比例为 0.5，字粒度模型的最优加噪比例为 0.4。噪音通过两种方式来添加，部分置 0，部分置随机数。在本次实验中，该比例设为 4∶1。

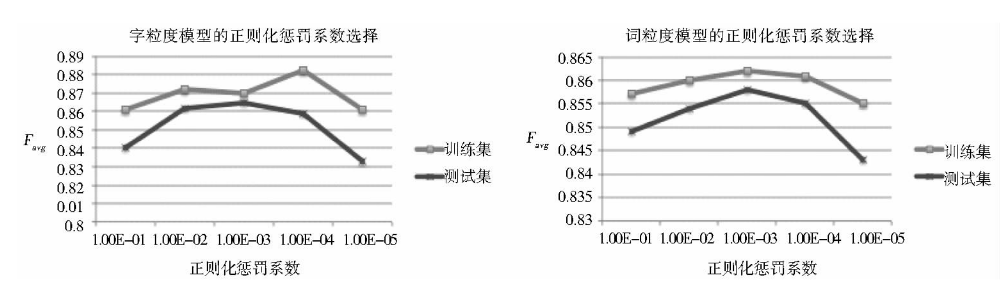
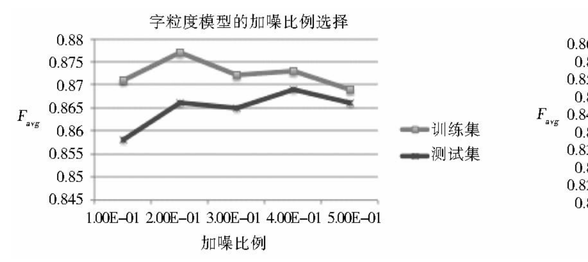


图 10　正则化惩罚系数选择



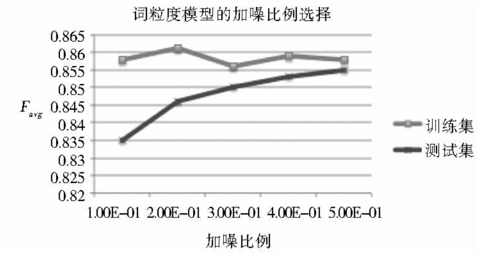


图 11　加噪比例选择

综合以上分析，字粒度模型在包含 4 层隐藏层，每层结点数设为 500，正则化惩罚系数 1e-3，加噪比例 40% 时达到最佳；词粒度模型在 3 层隐藏层，每层结点数 500，正则化惩罚系数 1e-3，加噪比例 50% 的情况下达到最佳。此时，算法在各情感倾向类别上的表现如表 4 所示：

表 4　字粒度和词粒度模型的 F 值对比

模型	正面	中性	负面	平均值
字粒度	0.871	0.862	0.865	0.866
词粒度	0.863	0.853	0.858	0.858

由此可知，基于字粒度的模型在各个情感倾向类别上都比基于词粒度的模型拥有更强的识别能力，说明对于微博语料而言，传统分词工具在切分工作中的确存在不够准确的问题，损失了部分信息，而字粒度的模型能够更好地保留微博中的信息，获取更加全面有效的特征。

图 12 是字粒度的详细实验结果，从图中可知，对正面的情感计算精确度最高，说明正向情感更容易判别，但是召回率较低，说明有一些正面情感丢失，而中性情感正相反，说明不少正面情感被分到了中性情感类，负面情感则介于中间，说明其特点不够显著，这也是本算法的特点，并没有依赖标注的情感体系或情感词库，但各项指标也都获得了较理想的结果。

6.4　对比分析

实验中设计了两组对比实验来验证 WE_SDAE 算法的合理性与有效性。根据前文的实验分析可知，字粒度模型的综合表现优于词粒度模型，因此，下文用于

Test Acc: 86.6%
Precision, Recall and F1-Score...

	precision	recall	f1-score	
正面	0.893	0.849	0.871	0.915
中性	0.835	0.891	0.862	0.808
负面	0.872	0.858	0.865	0.879

图 12　基于字向量的实验结果

对比的 WE_SDAE 算法都是基于字粒度实现的。

（1）WE_SDAE 与 WE_SVM 的对比分析。目前情感倾向分类使用最多的算法是 SVM，因此将在同样的数据集中训练 WE-SVM 算法，将其结果与 WE_SDAE 进行比较。

使用 SVM 工具包 LibSVM 来进行实验。数据预处理方式与基于字粒度的 WE_SDAE 算法完全相同，最终每条微博都会被表示成连续空间中的一个向量。经过调试优化，WE-SVM 在测试数据集上的效果与 WE_SDAE 进行了对比，结果如表 5 所示：

表 5　WE_SDAE 和 WE_SVM 算法的 F 值对比

算法	正面	中性	负面	平均值
WE_SDAE	0.871	0.862	0.865	0.866
WE_SVM	0.829	0.812	0.831	0.824

由此可知，WE_SDAE 算法的确优于传统的 WE-SVM，深层噪音自动编码器能够更好地提取微博的抽象特征，帮助分类器得到更为准确的情感倾向判定结果。

（2）WE_SDAE 与 VSM_SDAE 的对比分析。使用单词嵌入（Word Embedding）方法获取每条微博向量，其实目前在中文自然语言处理中，更为常见的方式是

使用向量空间模型(Vector Space Model)来处理微博数据,因此将把实验数据处理成 VSM 的形式,与 WE 方式进行对比。

实验首先将微博数据按照字粒度进行分割,然后去除那些出现次数小于 5 次的字,一共得到 5 236 个不同的字,则对应的微博向量长度也为 5 236。为了保证 VSM 方式和 WE 方式处理后的数据能够在同一个分类器中进行,需要对经 VSM 方式处理的微博向量进行降维。使用 Scikit-Learn 中的 PCA 算法将向量处理成 300 维,然后输入到分类器中进行训练,结果见表 6。其中,还给出了支持向量机(SVM)、朴素贝叶斯方法(N-Bayes)和集成分类(XgBoost)等算法的对比结果。

表 6 WE_SDAE 和 VSM_SDAE 算法的 F 值对比

算法	正面	中性	负面	平均值
WE_SDAE	0.871	0.862	0.865	0.866
VSM_SDAE	0.816	0.791	0.805	0.805
SVM	0.767	0.738	0.759	0.755
N-Bayes	0.711	0.731	0.695	0.712
XgBoost	0.756	0.787	0.743	0.762

由此可知,单词嵌入的向量获取方式的确优于传统的向量空间模型。一方面,单词嵌入在一定程度上解决了维数灾难问题,避免了降维操作;另一方面,它也能更好地挖掘出文本的语义与语境信息,有助于解决情感倾向分类问题。

7 结语

本文从三个方面提出了面向微博情感倾向分类的新思路:①考虑到通过传统空间向量模型获取的词向量既忽略了词与词之间的语义相关性,又缺失了语义分析中重要的上下文信息,使用单词嵌入的方式将微博中的每个词映射成连续空间中的一个向量,最大限度地保留微博文本自身的语义信息;②算法不再依赖任何人工标注的情绪知识体系,也不再拘泥于传统的机器学习模型,而是改进了深度学习中的自动编码器算法,借助其强大的无监督非线性学习能力,来完成微博特征的抽取与情感倾向的预测工作;③微博作为一种轻松社交媒体,用户在使用微博时的用语往往比较随意,存在大量的缩写。传统分词工具很可能无法准确地识别出这些信息,因而实验从单字和词语两个粒度来划分微博文本,尽可能地减少不必要的信息损失。

本文提出的算法会受到语料库的影响,语料库更大或者更专业会使字词编码更准确,有利于提高算法精度。同时,因为缺少情感词库的辅助,算法对情感词的把握不准,下一步可以考虑增加情感词的强化处理。同时,进一步的研究还包括结合句法分析和上下文语义关系等。

参考文献:

[1] TUMEY P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of annual meeting of the Association for Computational Linguistics. Stroudsburg PA: The Association for Computer Linguistics, 2002: 417-424.

[2] 任远, 巢文涵, 周庆, 等. 基于话题自适应的中文微博情感分析[J]. 计算机科学, 2013, 40(11):231-235.

[3] BARBOSA L, FENG J. Robust sentiment detection on Twitter from biased and noisy data[C]//Proceedings of 23rd international conference on computational linguistics. Cambridge: MIT Press, 2010:36-44.

[4] 庞磊,李寿山,周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程, 2012, 38(13):156-158.

[5] 潘明慧, 牛耘. 基于多线索混合词典的微博情绪识别[J]. 计算机技术与发展, 2014(9):28-32.

[6] 刘全超, 黄河燕, 冯冲. 基于多特征微博话题情感倾向性判定算法研究[J]. 中文信息学报, 2014, 28(4):123-131.

[7] BAKLIWAL A, FOSTER J, VAN DER PUIL J, et al. Sentiment analysis of political Tweets: towards an accurate classifier[C]//Proceedings of NAACL Workshop on language analysis in social media. Stroudsburg PA: The Association for Computer Linguistics, 2013:49-58.

[8] JOHAN B, ALBERTO P, HUINA M. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena[C]//Proceedings of 5th AAAI international conference on Weblogs and social media. Menlo Park, California: The AAAI Press, 2011:450-453.

[9] TAN C, LEE L, TANG J, et al. User-level sentiment analysis incorporating social networks[C]//Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2011:1397-1405.

[10] 刘志明, 刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1):1-4.

[11] 朱玺, 董喜双, 关毅, 等. 基于半监督学习的微博情感倾向性分析[J]. 山东大学学报:理学版, 2014, 49(11):37-42.

[12] 孙建旺, 吕学强, 张雷瀚. 基于词典与机器学习的中文微博情感分析研究[J]. 计算机应用与软件, 2014, 31(7):177-181.

[13] LIU N, ZHANG B, YAN J, et al. Text representation: from vector to tensor[C]//Proceedings of IEEE international conference on data mining. New Jersey: IEEE Press, 2005:725-728.

[14] KIM K, LEE J. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction[J]. Pattern recognition, 2014, 47(2):758-768.

[15] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation

of word representations in vector space[C]//Proceedings of international conference on learning representations. New York: ACM, 2013:1301－1309.
[16] MILOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 27th annual conference on neural information processing systems. Cambridge: MIT Press, 2013:3111－3119.
[17] MILOLOV T, YIH W, ZWEIG G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies. Stroudsburg: The Association for Computer Linguistics, 2013:746－751.
[18] ZHENG X, CHEN H, XU T. Deep learning for Chinese word segmentation and POS tagging[C]//Proceedings of the 2013 conference on Empirical methods in natural language processing. Stroudsburg: The Association for Computer Linguistics, 2013:647－657.
[19] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th international conference on machine learning. New York: ACM, 2008:1096－1103.
[20] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of machine learning research, 2015, 11(6):3371－3408.
[21] 桂斌, 杨小平, 朱建林,等. 基于意群划分的中文微博情感倾向分析研究[J]. 中文信息学报, 2015, 29(3):100－105.
[22] HASSAN S, HE Y, HARITH A. Semantic sentiment analysis of twitter[C]//Proceedings of the 11th international conference on the semantic Web. Berlin: Springer, 2012:508－524.
[23] PAK A, PAROUBEK P. Twitter as a corpus for sentiment analysis and opinion mining[C]//Seventh conference on international language resources & evaluation. Paris: European Language Resources Association, 2010: 1320－1326.
[24] SVETRIK V, LIAW A, TONG C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. Journal of chemical information & computer sciences, 2003, 43(6):1947－1958.
[25] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy layer-wise training of deep networks[C]//Proceedings of the 21th annual conference on neural information processing systems. Cambridge: MIT Press, 2007: 153－160.

作者贡献说明:
刘勘:论文的模型设计、架构设计,论文的修改和定稿;
袁蕴英:论文数据采集、实验和论文初稿。

Sentiment Classification for Micro-Blogs Based on Word Embedding

Liu Kan　Yuan Yunying
School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430074

Abstract: [**Purpose/significance**] Weibo has become an important platform for public emotional expression. Weibo's sentiment analysis plays an important role in public opinion analysis, user experience, and business opportunities. [**Method/process**] The sentiment orientation model named WE_SDAE proposed by this paper uses word embedding to transform a weibo into a dense low-dimensional vector and optimizes the simple auto-encoder into a deep denoise auto-encoder by appending a regularization term in the equation and adding noise during data pre-processing. Besides, the top-level classifier does the final sentimental classification. Considering the flexible term usage in the weibo, the sentiment orientation model is trained on character level and word level respectively. [**Result/conclusion**] The experimental results show that character-level model beats word-level model. In addition, comparative experiments show that WE_SDAE is better than traditional classifier SVM, Naive-Bayes, XgBoost, etc., and word embedding data preprocessing is better than traditional vector space model representation.

Keywords: sentiment analysis　classification　auto-encoder　Weibo